

# **Co-clustering of time-dependent data via the SIM**

Trinity College Dublin - Statistics Seminar Series

---

 Alessandro Casa

Joint work with: C. Bouveyron, E. Erosheva & G. Menardi

 Faculty of Economics and Management

Free University of Bozen-Bolzano

 alessandro.casa@unibz.it

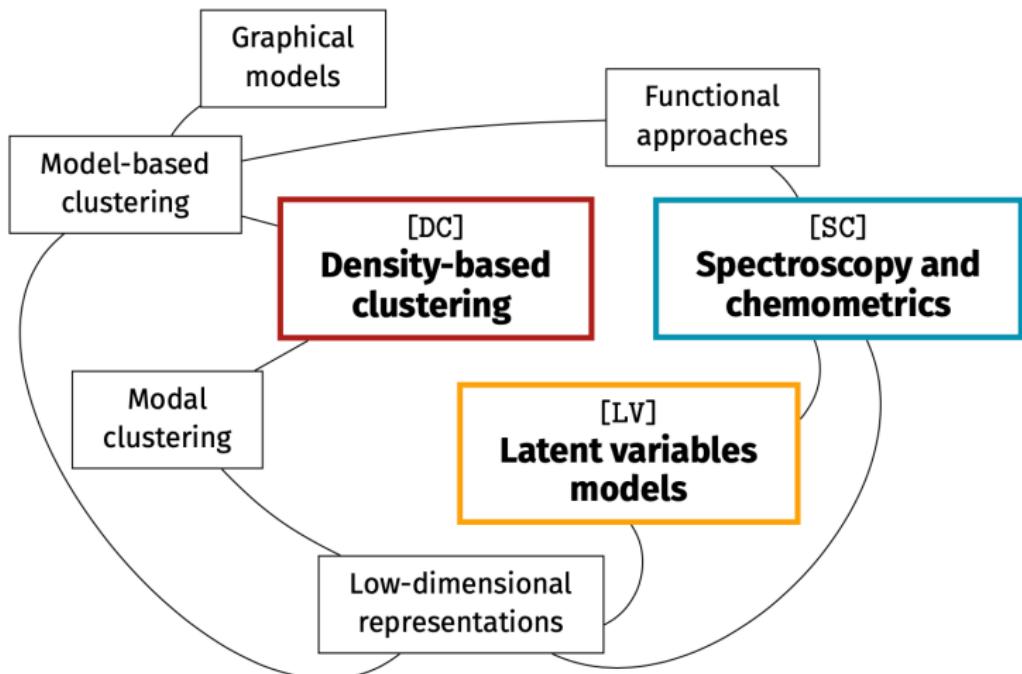
**unibz**

## ➤ Two words about me

- I am an Assistant Professor in Statistics at the Faculty of Economics and Management, Free University of Bozen-Bolzano
- Previously:
  - Postdoctoral researcher at UCD, School of Maths and Stats
  - Affiliated with Insight and Vistamilk SFI Research Centre
  - PI: Prof. Brendan Murphy
- Stone Age:
  - PhD in Statistics at University of Padova under the supervision of Prof. Giovanna Menardi
  - Lucky enough to spend some visiting periods in Nice, Cambridge and Perugia

## Research interests

My **current research** focuses on application-motivated problems where flexible modelling is needed to handle high-dimensional and complex structured data



## ➤ Framework - Time-dependent data

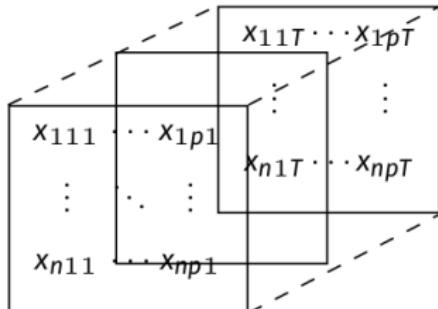
- Time-dependent data are everywhere
  - Stock market, economic indices, disease evolution...
- A taxonomy is tricky, we may distinguish between two poles
  - **Longitudinal data:** few observations, sparse and irregular measurements
  - **Functional data:** large number of observations, regularly sampled
- A lot of attention to the description of time evolutions and correlation between instants



What about possible heterogeneity among different trajectories?

## Framework

- Increasingly common multivariate time-dependent data can be arranged according to a three-way structure



- Three modes of the data introduce three different challenges
  - rows       $\longleftrightarrow$     heterogeneous units
  - columns     $\longleftrightarrow$     dependent variables
  - layers       $\longleftrightarrow$     correlated occasions
- Standard clustering methodologies fall short

### Aim

Extract information and unveil parsimonious patterns from such data

## ➤ Co-clustering in a nutshell

### Idea

Co-clustering tools may be helpful by summarizing the data into homogeneous blocks

- Given  $\mathbf{X} \in \mathbb{R}^{n \times p}$ , the Latent Block Model (LBM) is written as

$$p(\mathbf{X}; \Theta) = \sum_{\mathbf{z} \in Z} \sum_{\mathbf{w} \in W} \prod_{ik} \pi_k^{z_{ik}} \prod_{jl} \rho_l^{w_{jl}} \prod_{ijkl} p(x_{ij}; \theta_{kl})^{z_{ik} w_{jl}}$$

- $\Theta = (\pi_k, \rho_l, \theta_{kl})_{1 \leq k \leq K, 1 \leq l \leq L}$ , with  $K$  and  $L$  the number of row and column clusters
- $\mathbf{z} = (z_{ik})_{1 \leq i \leq n, 1 \leq k \leq K}$  and  $\mathbf{w} = (w_{jl})_{1 \leq j \leq p, 1 \leq l \leq L}$  denote the subject and variable cluster memberships
- $(x_{ij} | z_{ik} = 1, w_{jl} = 1) \sim p(\cdot; \theta_{kl})$
- Everything boils down to a proper specification of  $p(x_{ij}; \theta_{kl})$

## Time-dependent Latent Block Model

- In this framework  $\mathbf{X} = \{x_{ij}(\mathbf{t}_i)\}_{1 \leq i \leq n, 1 \leq j \leq p}$ ,  $\mathbf{t}_i = (t_{i,1}, \dots, t_{i,m_i})$   
→ each single cell is a curve
- We resort to the **Shape Invariant Model** (SIM) defined as

$$(x_{ij}(t) | z_{ik} = 1, w_{jl} = 1) = \alpha_{ij,1}^{kl} + e^{\alpha_{ij,2}^{kl}} m(t - \alpha_{ij,3}^{kl}; \beta_{kl}) + \varepsilon_{ij}(t)$$

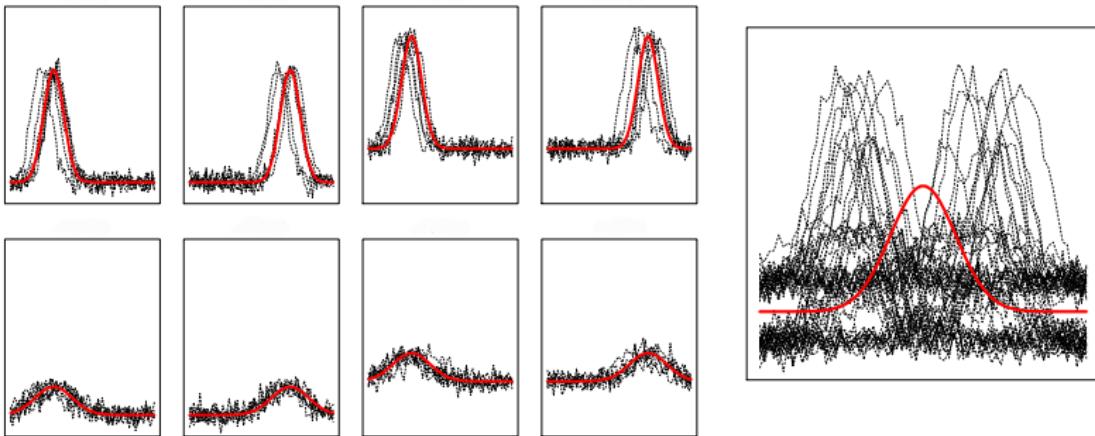
- $m(\cdot)$  block-specific mean shape function
- $\alpha_{ij}^{kl} = (\alpha_{ij,1}^{kl}, \alpha_{ij,2}^{kl}, \alpha_{ij,3}^{kl}) \sim \mathcal{N}(\mu_{kl}^\alpha, \Sigma_{kl}^\alpha)$  vector of cell and block-specific random parameters
- $\varepsilon_{ij}(t) \sim \mathcal{N}(0, \sigma_{\varepsilon,kl}^2)$

### Assumption

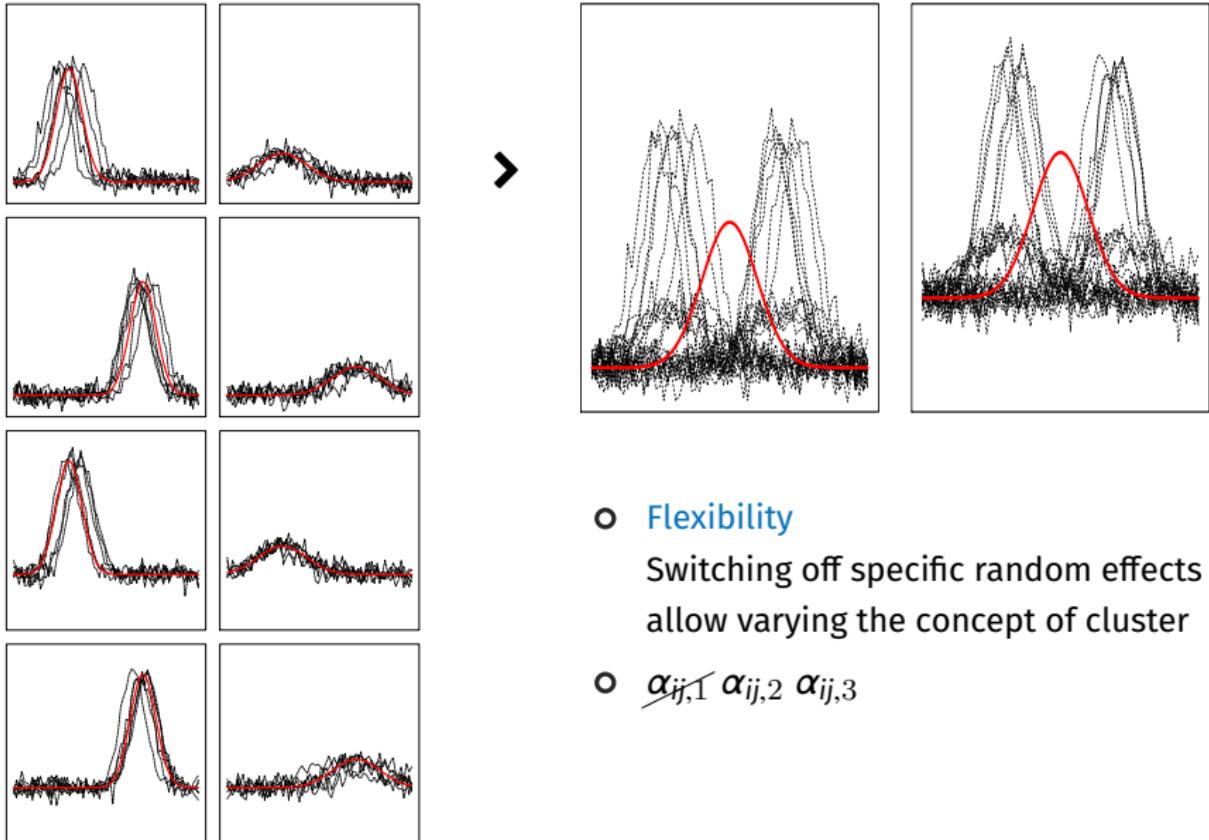
curves in a block arise as random transformation  
of a block-specific mean shape function

## ➤ Why SIM and co-clustering?

- It has been used to model functional and longitudinal data
- Blocks are characterized by the mean shape function but heterogeneity within them is appropriately modelled by  $\alpha_{ij}^{kl}$



## ➤ Why SIM and co-clustering?

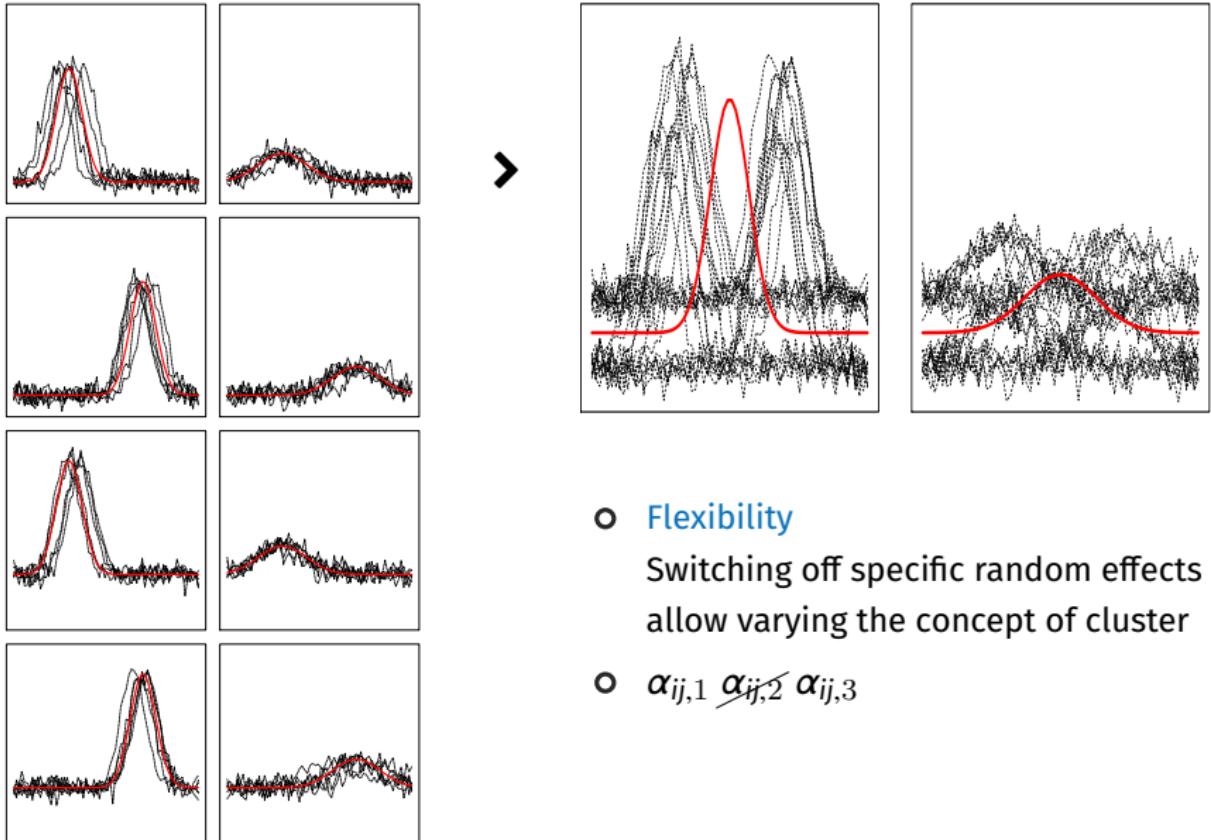


- **Flexibility**

Switching off specific random effects  
allow varying the concept of cluster

- $\alpha_{ij,1} \alpha_{ij,2} \alpha_{ij,3}$

## ➤ Why SIM and co-clustering?

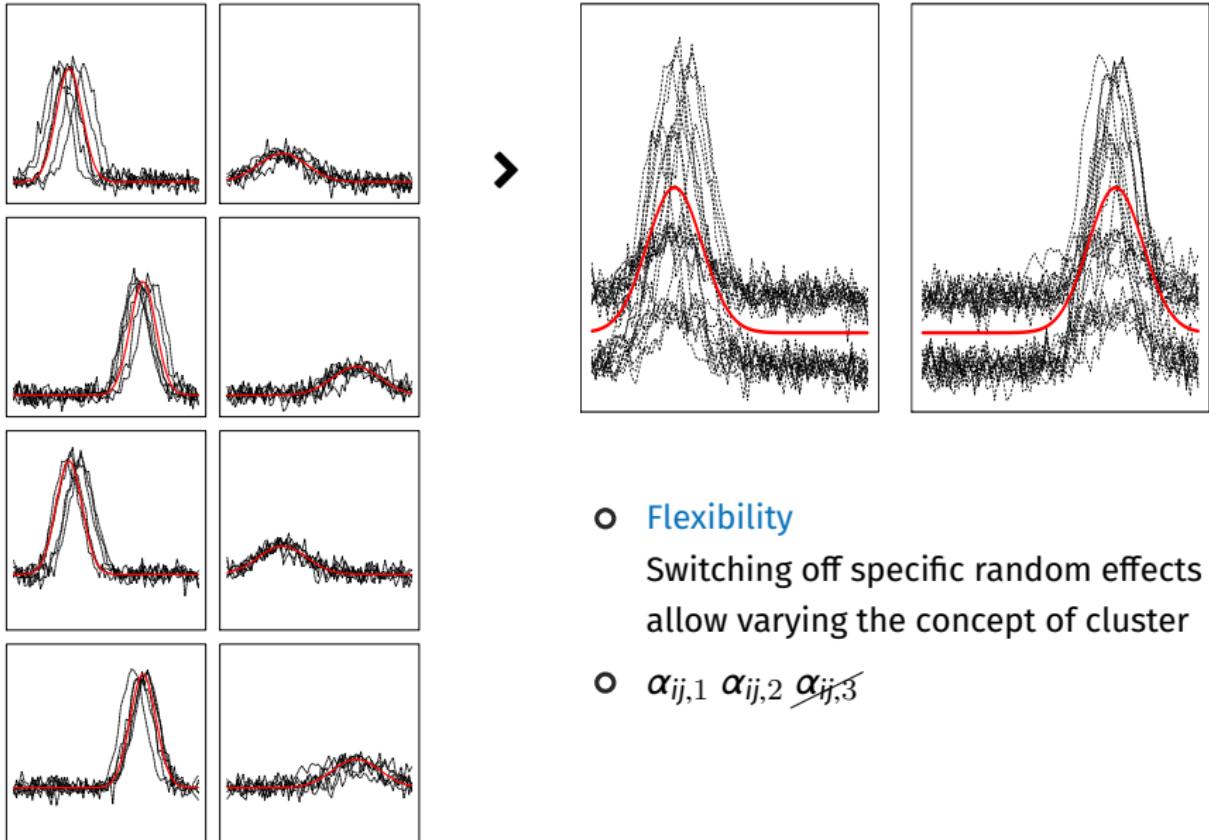


- **Flexibility**

Switching off specific random effects  
allow varying the concept of cluster

- $\alpha_{ij,1}$   ~~$\alpha_{ij,2}$~~   $\alpha_{ij,3}$

## Why SIM and co-clustering?



- **Flexibility**

Switching off specific random effects  
allow varying the concept of cluster

- $\alpha_{ij,1}$   $\alpha_{ij,2}$   $\cancel{\alpha_{ij,3}}$

## ➤ Model estimation - 1

- Maximization of the complete-data log-likelihood

$$\ell_c(\Theta, \mathbf{z}, \mathbf{w}) = \sum_{ik} z_{ik} \log \pi_k + \sum_{jl} w_{jl} \log \rho_l + \sum_{ijkl} z_{ik} w_{jl} \log p(x_{ij}; \theta_{kl})$$

- Double missing structure makes standard EM-algorithm computationally unfeasible in a co-clustering setting  
→ several modifications have been explored: CEM, SEM, VEM...
- **Additional problem:** no closed form expression for

$$p(x_{ij}; \theta_{kl}) = \int p(x_{ij} | \alpha_{ij}^{kl}; \theta_{kl}) p(\alpha_{ij}^{kl}; \theta_{kl}) d\alpha_{ij}^{kl}$$

since  $\alpha_{ij}^{kl}$  enters non-linearly in the model specification

## ➤ Model estimation - 2

- We propose the Marginalized SEM-Gibbs (M-SEM) algorithm
- At the  $h$ -th iteration we alternate these steps
  - Marg-step: obtain marginal distribution via MC integration

$$p(x_{ij}; \theta_{kl}^{(h-1)}) \simeq \frac{1}{M} \sum_{m=1}^M p(x_{ij} | \alpha_{ij}^{kl, (m)}; \theta_{kl}^{(h-1)}) ,$$

with  $\alpha_{ij}^{kl, (1)}, \dots, \alpha_{ij}^{kl, (M)}$  drawn from  $\mathcal{N}(\mu_{kl}^{\alpha, (h-1)}, \Sigma_{kl}^{\alpha, (h-1)})$

- SE-step: generate  $(\mathbf{z}^{(h)}, \mathbf{w}^{(h)})$  via Gibbs sampler
- M-step: Estimate  $\hat{\Theta}^{(h)}$  conditionally on  $(\mathbf{z}^{(h)}, \mathbf{w}^{(h)})$ 
  - Mixture proportions updated as usual
  - $\theta_{kl}^{(h)} = (\mu_{kl}^{\alpha, (h)}, \Sigma_{kl}^{\alpha, (h)}, \sigma_{\epsilon, kl}^{2, (h)}, \beta_{kl}^{(h)})$  estimated via approximate maximum likelihood approach for `nlme`

## ➤ Model selection

- Need to select  $K$  and  $L$ , the number of row and column clusters  
+ the random effect configuration (FFF, TFF, TTF...)
- In this framework we consider the ICL-BIC criterion

$$ICL = \ell_c(\hat{\Theta}, \hat{z}, \hat{w}) - \frac{K-1}{2} \log n - \frac{L-1}{2} \log d - \frac{KL\nu}{2} \log nd$$

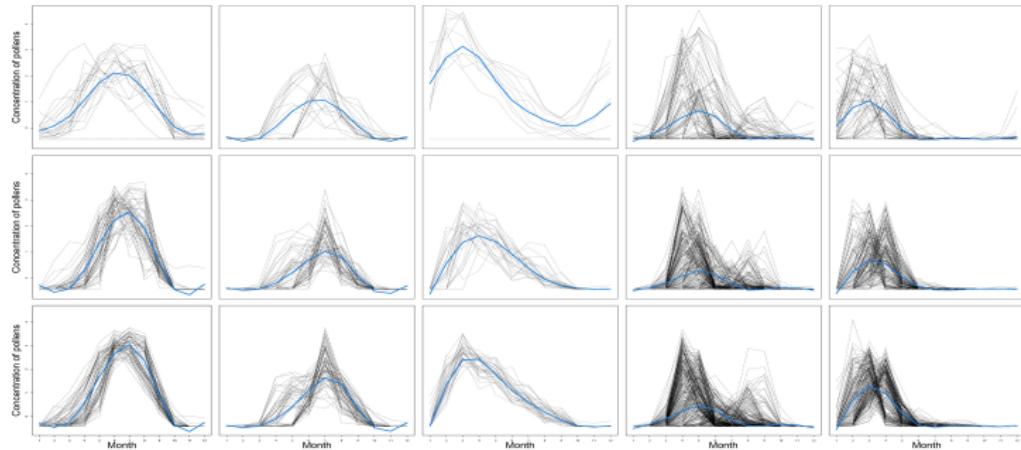
- Random effects and time-dependent data could make model selection more troublesome → bias towards overestimation

Incorporating prior knowledge and subject-matter considerations is highly beneficial

## Some results - Pollen data

- Monthly concentration of pollens in French cities in 2016  
 $p = 21$  pollens for  $n = 71$  cities over  $T = 12$  months
- Aim: identify homogeneous trends in pollen concentration over the year and across different geographic areas  
→ partition of both cities and pollens
- Searching for groups of pollens differentiating for either the period of exhibition or the time span they are present
  - ↓
  - Only models with y-axis shift (TFF) have been estimated,  
 $\alpha_{ij,2}$  and  $\alpha_{ij,3}$  are switched off

## Some results - Pollen data

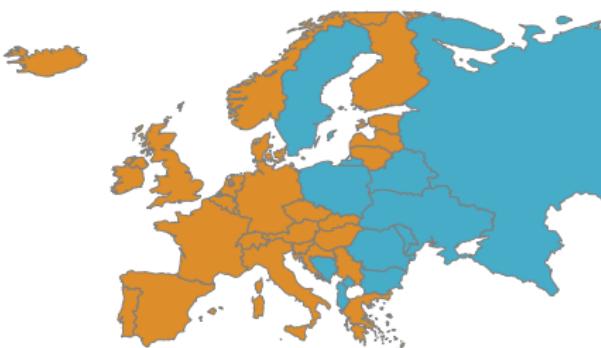


- Proper discrimination of pollens according to their seasonality + distinguish tree pollens from weed and grass ones
- Highlight a Mediterranean region

## Some results - COVID evolution

- Data on the first wave (1st March 2020 - 4th July 2020) for different European countries
- $n = 38$  countries,  $p = 4$  variables (daily cases, daily deaths, stringency index, government response index),  $T = 18$  weeks
- **Aim:** evaluate differences and similarities among countries and for different aspects of the pandemic
- Differently from pollens data, we have no reason to favour a specific random effects configuration
  - All the possible models have been estimated, with  $K = 1, \dots, 6$  and  $L = 1, 2, 3$
  - It entails different notions of similarity of virus evolution

## Some results - COVID evolution

Row groups	Column groups
 A choropleth map of Europe where countries are colored based on their COVID-19 status. Most countries are shaded in orange, while a few like Iceland, Norway, Sweden, and Russia are shaded in blue. The color coding likely represents the log percentage of cases per 1000 inhabitants.	<ol style="list-style-type: none"><li>1 log % of cases per 1000 inhabitants</li><li>2 log % of deaths per 1000 inhabitants</li><li>3 Stringency and gov response indexes</li></ol>

- Model TTT, with  $K = 2$  and  $L = 3$  is selected
- Partitions make sense, one is also geographical

## ➤ Concluding remarks

- The proposed method partitions three-way matrices in blocks of homogeneous curves
- Some relevant advantages with respect to the competitors
  - Interpretability of the results
  - Higher flexibility, different notions of cluster
  - Both longitudinal and functional data
- Directions and open questions
  - Possible alternative model selection approaches
  - Different specification for the mean shape function
  - Bayesian estimation strategies

## Some references

Casa, A., Bouveyron, C., Erosheva, E. & Menardi, G. (2021).  
Co-clustering of time-dependent data via the Shape Invariant Model.  
*J Classif*, doi.org/10.1007/s00357-021-09402-8

### Other relevant references

- Bouveyron, C., Bozzi, L., Jacques, J. & Jollois, F.X. (2018). The functional latent block model for the co-clustering of electricity consumption curves.  
*J R Stat Soc C*, 67(4), 897-915.
- Lindstrom, M.J. (1995). Self-modelling with random shift and scale parameters and a free-knot spline shape function. *Stat Med*, 14(18), 2009-2021.
- Pinheiro, J. & Bates, D. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J Comput Graph Stat*, 4(1), 12-35.
- Telesca, D. & Inoue, L.Y.T. (2008). Bayesian hierarchical curve registration.  
*J Am Stat Assoc*, 103(481), 328-339.